Case Study 2:
Who Plays Video Games?

By:

Himanshu Makhija (A09845605) - Math/Computer Science, B.S.

Amey Paranjape (A53218045) - Electrical & Comp. Engineering, M.S.

Hanna Goldman (A11436767) - Math/Computer Science, B.S.

Shagun Gupta (A91068956) - Bioengineering: Bioinformatics, B.S.

(UCSD MATH 189/289C Winter 2017)

Table of Contents

Introduction

Analysis

Conclusion

Theory

Introduction

Objective

The objective of this study is to investigate the responses of the participants of the study with the intention of providing useful information about the students to build a computer lab.

Survey Methodology

In order to help statistics students at the University of California, Berkeley in 1994, a committee of both faculty and students was formed to help construct a series of computer labs as supplemental instruction. Specifically, the labs were designed for the undergraduate students enrolled in a lower division statistics course for intending business majors (Intro. to Probability and Statistics).

As such, the committee designed a survey to understand the time the students spent playing video games, what aspects of the games the enjoyed, and what they didn't enjoy. This survey was conducted by the advanced statistics students.

The class met Mondays, Wednesdays and Fridays from 1-2PM in a lecture hall that sat up to 400 students. In addition to this lecture, the students were also enrolled in a small, one-hour discussion/lab that met on Tuesdays and Thursdays. To ensure accuracy of results, the survey was conducted only after the second exam of the semester had been given. The week following this exam, the survey was given and data collectors visited both of the discussion sections the same week. This worked well because the exams were given back during this discussion meeting as well, allowing the collectors to count each survey given.

By Friday, the students that could not be reached through discussion were located during the lecture and given their survey. Finally, to encourage accuracy and honesty, the collectors informed the student of the purpose of the survey and guaranteed anonymity.

The population size, **N**, is **314** (the total size of the class). However, not all students were chosen to survey and not all of those chosen responded. Of the original N students, 95 were chosen, but of those only **91** returned their surveys (**n**). The probability model chosen to collect was simple random, without replacement, making the dataset non-i.i.d.

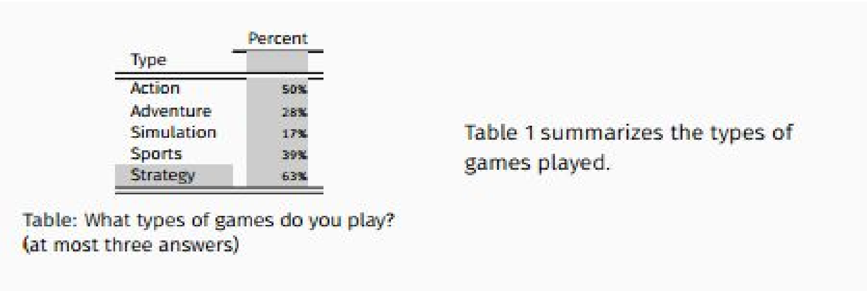
Description of Dataset

1) Time: (# hrs. played in week prior to survey) → numerical, discrete
2) Like to play: [1-5], 1=never, 2=very, 3=some, 4=not really, 5=nope → categorical, nominal

3) <u>Where</u> <u>play</u>: [1-6], 1=arcade, 2=home sys, 3=home comp, 4=arcade & (home OR comp), 5=home (comp & sys), 6=all → categorical, nominal
4) <u>How often</u>: [1-4], 1=daily, 2=weekly, 3=monthly, 4=semesterly → categorical, nominal
5) <u>Play if busy</u>: 1=yes, 0=no → categorical, nominal
6) <u>Play educational</u>: 1=yes, 0=no → categorical, nominal
7) <u>Sex</u>: 1=male, 0=female → categorical, nominal
8) <u>Age</u>: (# of years) → numerical, discrete
9) <u>Computer @ home</u>: 1=yes, 0=no → categorical, nominal
10) <u>Hate math</u>: 1=yes, 0=no → categorical, nominal
11) <u>Work</u>: (# hrs. worked week prior to survey) → numerical, discrete
12) <u>Own PC</u>: 1=yes, 0=no
13) <u>PC has CD-ROM</u>: 1=yes, 0=no
14) <u>Have email</u>: 1=yes, 0=no
15) <u>Grade expected</u>: 4=A, 3=B, 2=C, 1=D, 0=F → categorical, ordinal

Note: The questions that were not answered were coded as ‘99’, which when used was changed to ‘NA’ to remove its effect on our plots and results.

While the above dataset had only one response for each variable, the second part of the survey allowed respondents to answer in 3 categories.

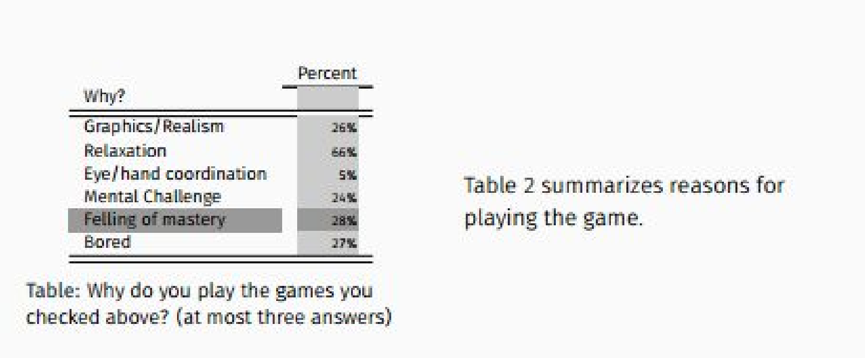


Type	Percent
Action	50%
Adventure	28%
Simulation	17%
Sports	39%
Strategy	63%

Table: What types of games do you play? (at most three answers)

Table 1 summarizes the types of games played.

<u>Plot 1.1</u>: “What types of games do you play?” was not answered by the full sample, since those that have never played any video games were asked to intentionally leave it blank

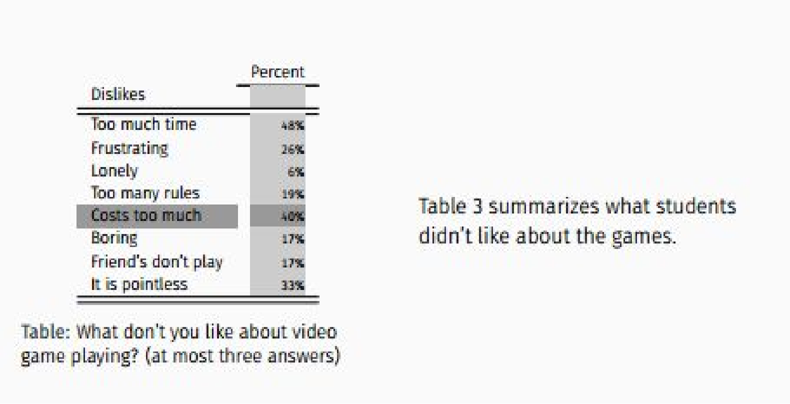


Why?	Percent
Graphics/Realism	26%
Relaxation	66%
Eye/hand coordination	5%
Mental Challenge	24%
Felling of mastery	28%
Bored	27%

Table: Why do you play the games you checked above? (at most three answers)

Table 2 summarizes reasons for playing the game.

<u>Plot 1.2</u>: “Why do you play the games you check above?” was only answered by the students Who answered the question above

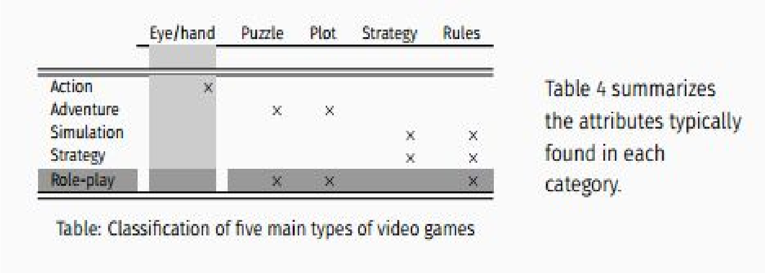


Dislikes	Percent
Too much time	48%
Frustrating	26%
Lonely	6%
Too many rules	19%
Costs too much	40%
Boring	17%
Friend's don't play	17%
It is pointless	33%

Table: What don't you like about video game playing? (at most three answers)

Table 3 summarizes what students didn't like about the games.

Plot 1.3: "What don't you like about video game playing?" was answered by all survey respondents (n = 91)



	Eye/hand	Puzzle	Plot	Strategy	Rules
Action	×				
Adventure		×	×		
Simulation				×	×
Strategy				×	×
Role-play		×	×		×

Table: Classification of five main types of video games

Table 4 summarizes the attributes typically found in each category.

Plot 1.4: "Classification of the five main types of video games" shows the focus of the video game

Note: Video games can also be classified according to the following for the above table (1.4), but in terms of console the genre is played on;

- Arcade: Eye/hand
- Console: Eye/hand, Puzzle, Plot, Strategy, Rules
- PC: All attributes

Further Research of Video Games in Education

According to Kurt Squire from MIT, the video game industry is pervasive in our country which has been more or less ignored by the education system. Although some have researched the correlation with the ultimate goal of increasing the students' commitment to learn by introducing video game elements into the learning environment. They believe these video game elements generate a longing to play the video game, which should be considered when the ultimate goal for our case study is to build a lab that can help students with learning statistics effectively. Some factors include:

- Clear goals that students find meaningful
- Multiple goal structures for clarity
- Giving feedback on progress
- Different difficulty levels to introduce learners and adjust experienced gamers

- Randomness of plot, to keep attention
- And finally, “emotionally appealing fantasy and metaphor that is related to video game skills”

In video games, players pursue their own goals and are tested to the extent of their abilities at which point they are provided personalized feedback, whereas in classrooms, students are forced to conform to the pace of the group and may receive impersonal feedback. While not part of our dataset findings, these factors will prove important in designing this lab.

Hypothesis

Playing video games an intermediate amount of time leads to better grades. Playing too little or too often can lead to la decrease in grades or grade expectancy. In terms of the amount that video games are played, people who work more or don’t have the resources to play will be less likely to play video games while people who hate math will be more likely to play resources. We think designing a lab that meets only once a week for a few hours will encourage students to attend because it will provide the resources and time in order to teach in a stimulating yet fun manner.

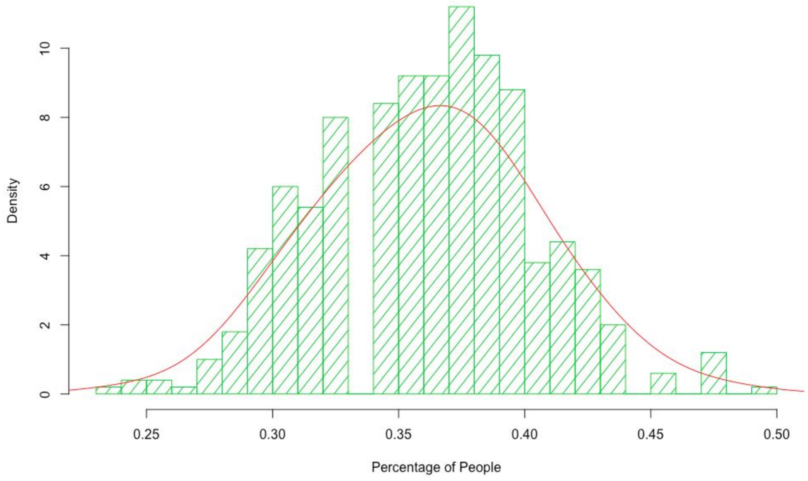
Analysis

Fraction of Students Who Play Video Games

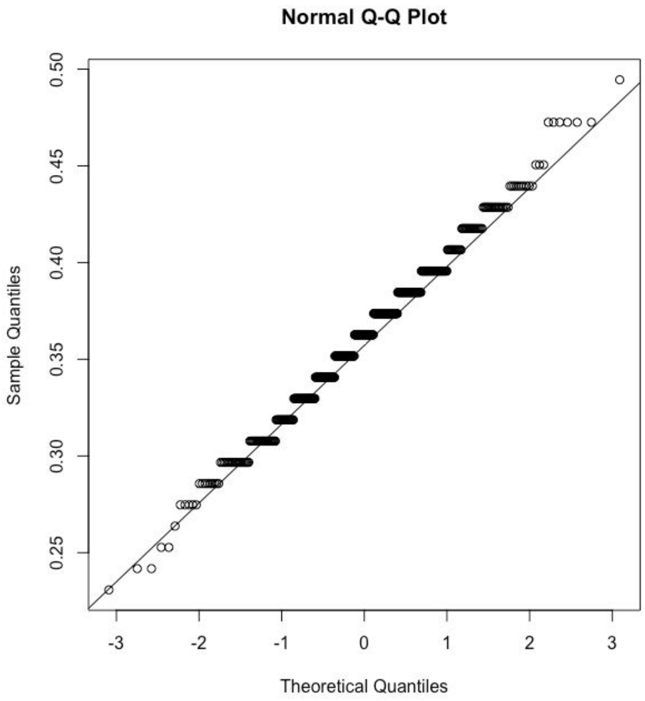
By observing the numbers of hours played by each student in the week prior to the survey, the data was split into whether the students had played (> 0 hours) or hadn't played(= 0 hours).
The data shows that in the sample, 34 out of 91 students played video games in the week prior to the survey.
This results in a estimate for the population parameter of ス = 34/91 = 0.3736264.
The standard error $\frac{\sqrt{\frac{34}{91}(1-\frac{34}{91})}}{\sqrt{91-1}} \times \frac{\sqrt{314-91}}{\sqrt{314}}$ =0.04297361 was then calculated in order to form a confidence interval for the population parameter.

Hence the 95% confidence interval is (0.3736264-(1.96*0.04297361), 0.3736264+(1.96*0.04297361)) = (0.2893981, 0.4578547)

A new population of size 314 was then created based on the sample and this bootstrap population was used to find a potential probability distribution for the population.
We took 500 bootstrap samples of size 91 from the bootstrap population, in order to make a reasonable simulation of a probability distribution of the bootstrap average.



Plot 2.1: Bootstrapped population Histogram
The data appears to be unimodal with zero to little skew. This suggests a normal distribution.



Plot 2.2: Normal Q-Q Plot to check normality of bootstrapped population samples

The circles all lie quite close to the line and it appears that the data follows a normal distribution. It is has light tails and the data is a little farther away from the line towards the tails but are not significant enough to be non-normal.

To further validate that the bootstrap population follows a normal distribution, the kurtosis and skewness for the 500 bootstrap sample averages were compared to a simulated distribution of skewness and kurtosis for a samples of size 500 from a normal distribution
Bootstrap skewness = 0.01291549
Bootstrap kurtosis = 3.000931

Normal skewness = 0.008031122
Normal kurtosis = 2.936867

Interpretation: The skewness of the simulated data is 0.01291549. This concludes that the data is close to bell shape but slightly skewed to the right. The computed kurtosis is 3.000931, which means the data is mesokurtic(coming from normal distribution). When compared to a random sample from a normal distribution, the deviation of skewness and kurtosis from what is considered normal were very similar. Therefore we can assume that our data comes from a normal distribution.
The mean of the bootstrap populations is 0.3614505.
The 95% Confidence Interval for the bootstrap data is (0.2857143, 0.4395604)
The mean and confidence interval of the bootstrap population are very close to the estimate for the population parameter and the confidence interval that we calculated for the sample.

Amount of Time Spent Playing Video Games vs Frequency of Play

In order to compare the frequency of play with the number of hours played, the data was first split into four categories based on how often the students recorded that they played: daily, weekly, monthly, or semesterly.
Next the hours played for each category of the sample was observed.

Daily

Min	Median	Mean	Max	Standard Deviation
0.000	2.000	4.444	14.000	5.570258

Weekly

Min	Median	Mean	Max	Standard Deviation
0.000	2.000	2.539	30.000	5.499046

Monthly

Min	Median	Mean	Max	Standard Deviation
0.000	0.000	0.05556	0.50000	0.1616904

Semesterly

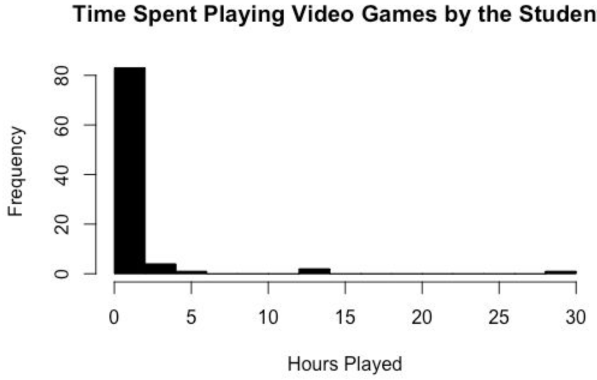
Min	Median	Mean	Max	Standard Deviation
0.000	0.000	0.04348	1.00000	0.2085144

It can be observed that on average, students who indicated that they played daily also tended to play more hours in the previous week. The section of students who played monthly had an outlier of a student who played for 30 hours in the previous week and still their mean was lower than the daily students.

The test was given in the week prior to the survey and this might have affected the number of students who played video games that week. Students who played video games on a daily basis are more likely to have kept up with this habit of playing even though they had a test, while students who played only weekly or monthly would be less likely to play because they might have been busier or less likely to be distracted by games. This might attribute to the fact that students who played daily had a higher mean than the rest of the categories of students.

Average Amount of Time Spent Playing Video Games

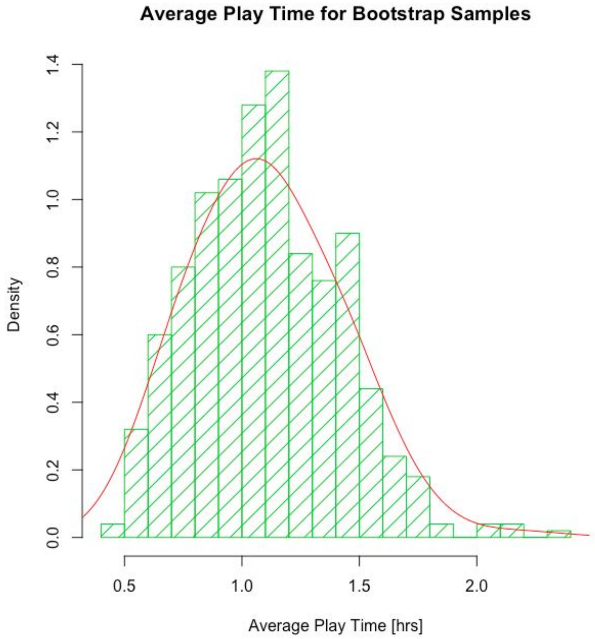
First, the sample distribution of the number of hours students played was observed.
From the histogram of the time spent playing video games by the students in the sample, it is clear that the sample distribution is extremely skewed.



Plot 2.3: Hours played plotted against the amount of players

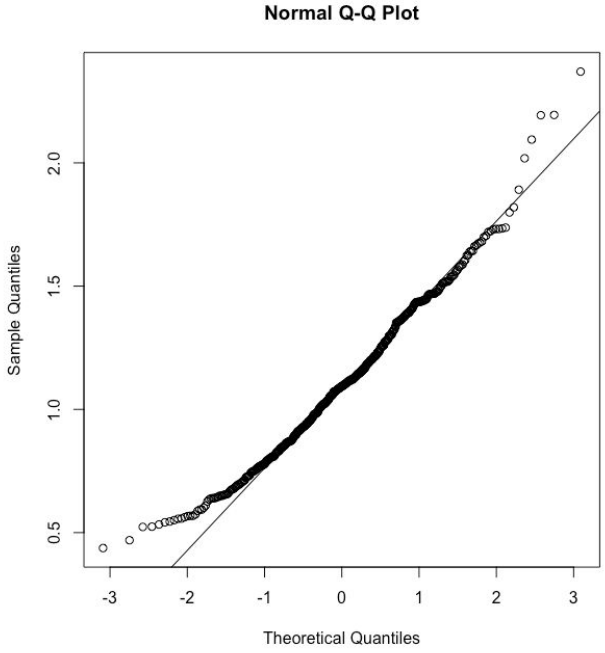
This observation raises a question of whether the probability distribution of the sample average follows a normal curve. Without knowing the distribution that the population follows, we can not make estimates about the population. Because simple random sampling is a fair way to select a sample, it is reasonable to generalize the results from the sample back to the population.

Bootstrap sampling was decided as the best option to make estimates for the data. A new population of size 314 was then created based on the sample and this bootstrap population was used to find a potential probability distribution for the population. We took 500 bootstrap samples of size 91 from the bootstrap population, in order to make a reasonable simulation of a probability distribution of the bootstrap average.



Plot 2.4: Average play from bootstrapped population over 500 samples

Although this data appears to have almost a normal distribution, it is slightly skewed right with a longer right tail.



Plot 2.5: Q-Q test for normality of distribution from bootstrap

The QQ-plot for this data shows a similar pattern. Although it sticks to a normal distribution in the center of the data, the tails are further from the normal line with a heavier right tail.

Shapiro-Wilk normality test

```
data:  boot.mean
W = 0.98256, p-value = 1.056e-05
```

As shown above, the Shapiro Test tests the NULL hypothesis that the samples came from a Normal distribution. Since the p-value <= 0.05, then we reject the NULL hypothesis that the samples came from a Normal distribution.

We also compared the kurtosis and skewness for the 500 bootstrap sample averages to a simulated distribution of skewness and kurtosis for a sample of this size

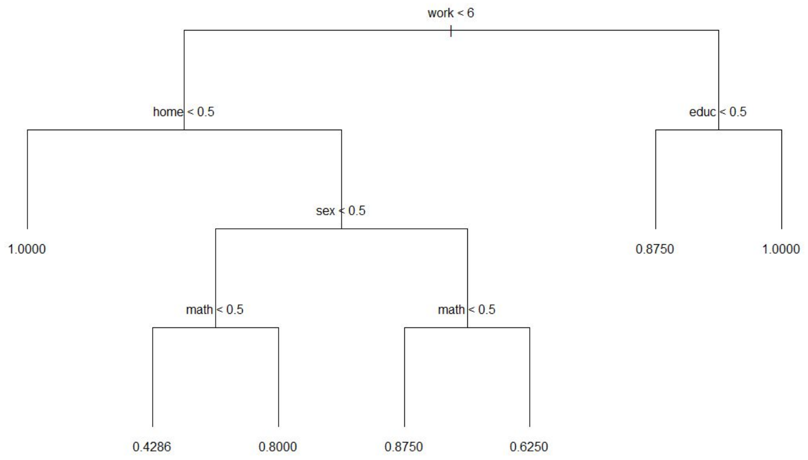
Bootstrap skewness = 0.4678843
Bootstrap kurtosis = 3.343751

Normal skewness = 0.0441108
Normal kurtosis = 3.008416

The point estimate of this bootstrap sample is 1.105897.
The 95% Confidence Interval
(0.5681319, 1.7271154)

Why Students Like/Dislike Video Games

After making a subset of the population wherein students were sorted based on their answer for whether they like playing/ not playing/ haven't ever played video games, the population was analyzed. Using the tree package for classification:



Plot 2.7: Tree of the factors affecting the liking/disliking of video game playing

```
tree(formula = dis_like ~ educ + sex + age + home + math + work +
    own + cdrom + grade, data = data)
```

```
Variables actually used in tree construction:
[1] "work" "home" "sex"  "math" "educ"
Number of terminal nodes:  7
Residual mean deviance:  0.1058 = 6.139 / 58
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.8750  0.0000  0.0000  0.0000  0.1250  0.5714
```

<u>Using regression trees:</u>

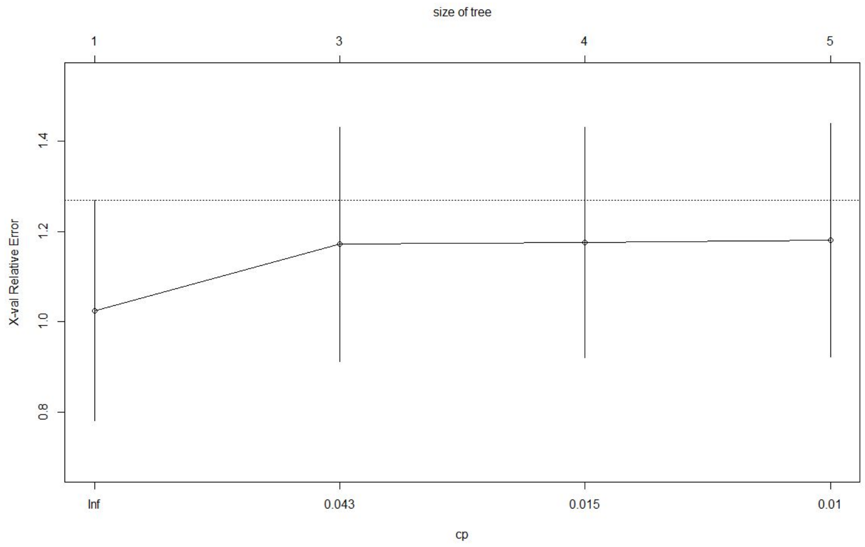
Regression tree:

Why we used regression tree? Our data interacts in a number of non linear ways with a combination of numerical and categorical data. A regression tree helps us deal with this type of data by partitioning this data into small portions which are analyzed separately as smaller populations. This is called recursive partitioning and is another way of dealing with nonlinear regression.

Another reason for this was that rpart package allowed a node by node analysis proving the number of observation at each branch and leaf (end nodes of the tree).

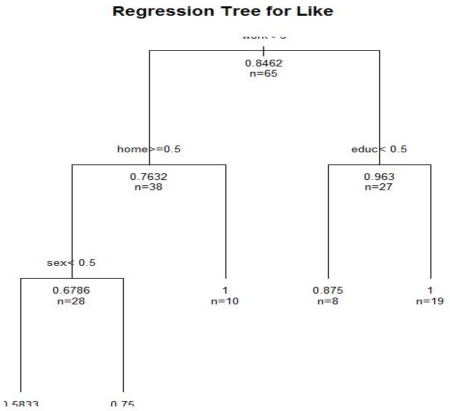
```
rpart(formula = dis_like ~ educ + sex + age + home + math + work +
    own + cdrom + grade, data = data)

Variables actually used in tree construction:
[1] educ home sex  work

Root node error: 8.4615/65 = 0.13018

n= 65  (our data has 65 entries that we graph)
> plotcp(fit) # visualize cross-validation results
```

The horizontal line below indicates the highest cross-validated error less than the minimum cross-validated error plus the standard deviation of the error at that tree. **Cross-validation**, sometimes called rotation estimation, is a model **validation** technique for assessing how the results of a statistical **analysis** will generalize to an independent data set.



Plot 2.8: Graph for further analysis of the suitability of the predicted curve on the observed values based on the relative error with increasing splits of the regression tree.

```
> summary(fit) # detailed summary of splits
Call:
rpart(formula = dis_like ~ educ + sex + age + home + math + work +
    own + cdrom + grade, data = data)
  n= 65
          CP           nsplit   rel error   xerror      xstd
1 0.08222102      0   1.0000000 1.024533 0.2437357
2 0.02251082      2   0.8355580 1.171241 0.2588802
3 0.01039562      3   0.8130471 1.175746 0.2545887
4 0.01000000      4   0.8026515 1.180639 0.2574299
Variable importance
 home  work  educ   own   sex grade   age
   34    28    12    10     9     3     3
```



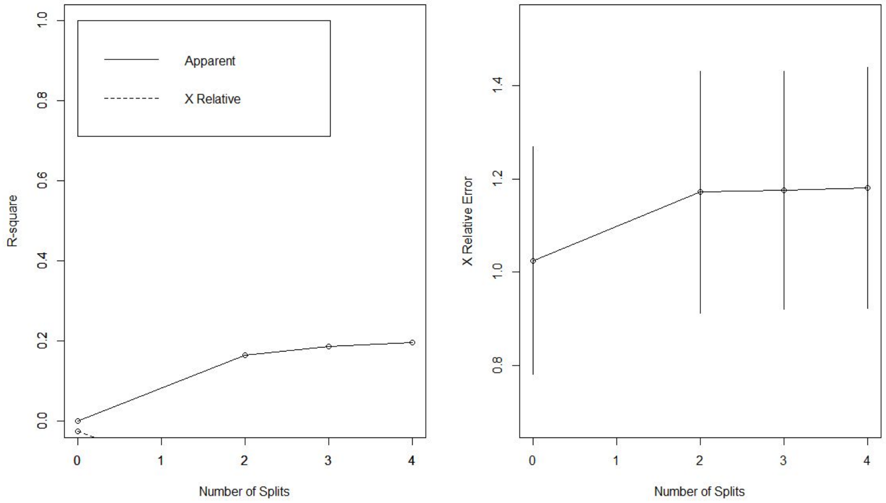
Plot 2.9: **The regression tree for all the factors affecting the liking or dislike of playing video games.**

The node by node information of the regression tree is given below:

```
Node number 1: 65 observations,    complexity param=0.08222102
  mean=0.8461538, MSE=0.1301775
  left son=2 (38 obs) right son=3 (27 obs)
  Primary splits:
      work < 6    to the left,  improve=0.07447280, (0 missing)
      educ < 0.5  to the left,  improve=0.06887052, (0 missing)
      home < 0.5  to the right, improve=0.05936920, (0 missing)
      math < 0.5  to the right, improve=0.04518590, (0 missing)
      own  < 0.5  to the right, improve=0.02456551, (0 missing)
  Surrogate splits:
      educ < 0.5  to the left,  agree=0.662, adj=0.185, (0 split)
      age  < 23.5 to the left,  agree=0.631, adj=0.111, (0 split)
      own  < 0.5  to the right, agree=0.631, adj=0.111, (0 split)

Node number 2: 38 observations,    complexity param=0.08222102
  mean=0.7631579, MSE=0.1807479
  left son=4 (28 obs) right son=5 (10 obs)
  Primary splits:
      home  < 0.5  to the right, improve=0.11083740, (0 missing)
      educ  < 0.5  to the left,  improve=0.02850757, (0 missing)
      sex   < 0.5  to the left,  improve=0.02303204, (0 missing)
      own   < 0.5  to the right, improve=0.01103519, (0 missing)
      cdrom < 0.5  to the left,  improve=0.01103519, (0 missing)
  Surrogate splits:
      own   < 0.5  to the right, agree=0.763, adj=0.1, (0 split)
      grade < 2.5  to the right, agree=0.763, adj=0.1, (0 split)

Node number 3: 27 observations,    complexity param=0.01039562
  mean=0.962963, MSE=0.03566529
  left son=6 (8 obs) right son=7 (19 obs)
  Primary splits:
      educ  < 0.5  to the left,  improve=0.09134615, (0 missing)
      age   < 19.5 to the right, improve=0.07692308, (0 missing)
      work  < 11   to the left,  improve=0.07692308, (0 missing)
      own   < 0.5  to the right, improve=0.02262443, (0 missing)
      grade < 3.5  to the left,  improve=0.01923077, (0 missing)
  Surrogate splits:
      own < 0.5  to the left,  agree=0.778, adj=0.25, (0 split)

Node number 4: 28 observations,    complexity param=0.02251082
```

```
  mean=0.6785714, MSE=0.2181122 
  left son=8 (12 obs) right son=9 (16 obs)
  Primary splits:
      sex   < 0.5  to the left,  improve=0.0311890800, (0 missing)
      educ  < 0.5  to the left,  improve=0.0093567250, (0 missing)
      age   < 19.5 to the left,  improve=0.0019493180, (0 missing)
      math  < 0.5  to the left,  improve=0.0007497376, (0 missing)
      grade < 3.5  to the right, improve=0.0004873294, (0 missing)
  Surrogate splits:
      educ < 0.5  to the right, agree=0.714, adj=0.333, (0 split)
      own  < 0.5  to the left,  agree=0.679, adj=0.250, (0 split)

Node number 5: 10 observations
  mean=1, MSE=0 

Node number 6: 8 observations
  mean=0.875, MSE=0.109375 

Node number 7: 19 observations
  mean=1, MSE=0 

Node number 8: 12 observations
  mean=0.5833333, MSE=0.2430556 

Node number 9: 16 observations
  mean=0.75, MSE=0.1875 
```

#Visualisation of the cross validation result.
Higher the R-squared value, better is the regression fit. Thus as number of splits increase, we get a better fit of the graph (and then it tapers off, i.e. no increase). Thus at lower number of splits, a lower R-squared value implies that our data is very hard to predict which is true for our data type.
X relative error increases and tapers of with increase in number of splits. This implies that as number of splits increase in our regression tree, we get better at predicting our values.



Plot 2.10: Further analysis of the prediction curve proposed in the regression tree for the observed entries, based on the trend of the R-squared values and the X relative error with increasing number of splits.

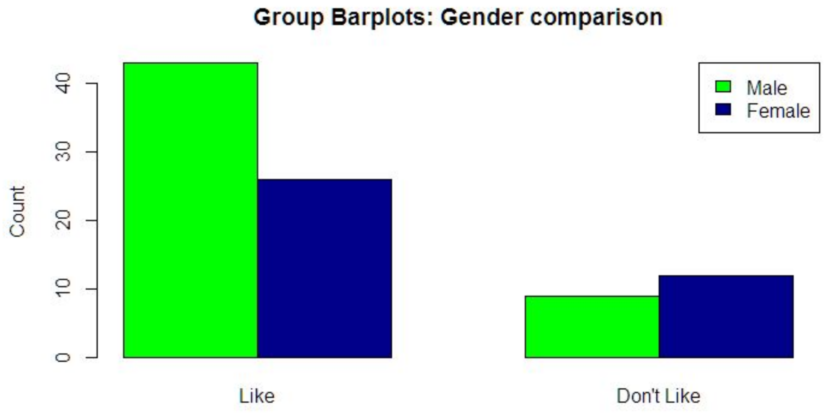
The graph implied that the greatest trend observed in students who liked playing video games, was that they worked for less than 6 hours in the week prior to the survey, had a greater chance of having a computer at home and were more likely to be male. It also showed that out of those who played for more than six hours a week, more people enjoy playing for educational purposes. Their responses are presented in the summary of the graph per node and were analyzed based on which branch had a greater number of entries and what the classifying factor was at that split.

This coupled with the tables above show that the major reasons for liking to play could be attributed to relaxation, and a feeling of mastery attained (through educational games, if a greater number of hours in invested in it). Thus the video game should be educational to the extent that it provides players with obvious skills in the recourse of being played.

The major reasons cited for disliking games were that it cost too much, was pointless or took too much time. Looking at the trend for those who like playing video games, it can be safely said that they would contribute to the factor of the game costing too much (though they own computers so the cost wouldn't be that big of an issue with them). The game taking too much time and being pointless (not educational) would also fit the trend observed in people who like video games.

Breaking Down the Variables

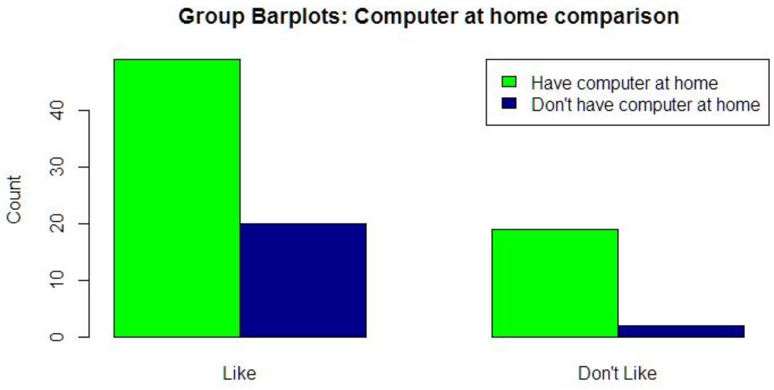
We compare the population based on preferences for video games versus gender, finding playing educational and owning a computer. In particular, we are searching for glaring differences in our sample and trying to determine strong attributes contributing to an individual’s preferences to video games.



<u>Plot 2.11</u>:

	Male	Female
Like	62.31%	37.68%
Dislike	42.85%	57.14%

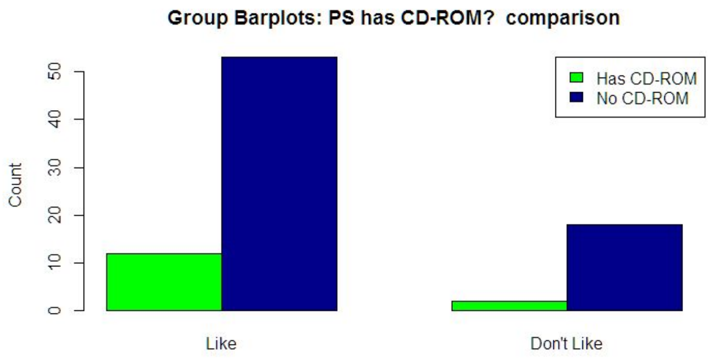
Comparing students who like playing video games and don’t like playing video games, based on the gender of the students, it can be observed that students who like video games are dominantly males and those who don’t like playing games are majorly females.



<u>Plot 2.12</u>:

	Own PC	Do not own PC
Like	71.01%	28.98%
Dislike	90.47%	9.52%

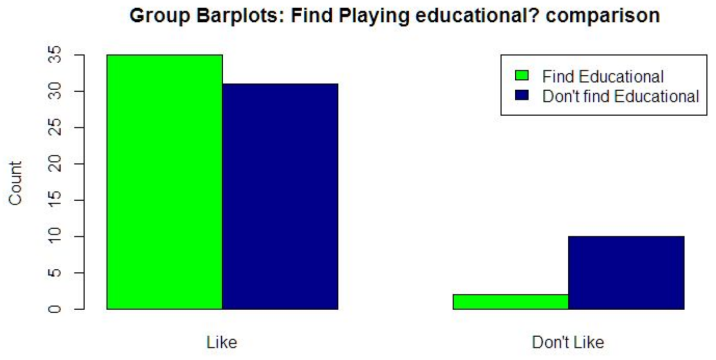
Comparing students who like playing video games and don't like playing video games, based on whether or not students own a PC, it can be observed that in students who dislike video games, there is a high percentage of students who own PC. Though this observation is not significant enough to infer anything.



Plot 2.13:

	CD-ROM	No CD-ROM
Like	18.46%	81.53%
Dislike	10%	90%

Comparing students who like playing video games and don't like playing video games, based on whether or not students' PS has a CD-ROM, it can be easily observed that majority of students who like playing games do not have a CD-ROM.

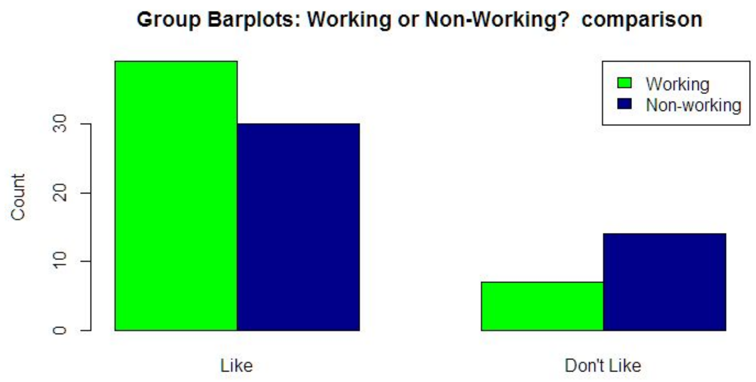


Plot 2.14:

	Find Educational	Don't find educational
Like	53.03%	46.96%

Dislike	16.66%	83.33%

Comparing students who like playing video games and don’t like playing video games, based on whether or not they find it Educational, it can be observed that there is no such concrete reason why people actually like playing video games as the difference between the percentage of people who like playing and also find it educational is almost equivalent to the percentage of people who like playing it and do not find it educational. Also, it can be observed that people who do not like playing video games,majorly do not find them educational.



<u>Plot 2.15</u>:

	Working	Non-working
Like	56.52%	43.47%
Dislike	33.33%	66.66%

Comparing students who like playing video games and don’t like playing video games, based on whether or not they are working for it or not, it can be observed that there is no such concrete observation which can be inferred from the above bar plots.

As we can see, the differences between liking and disliking depending on a particular attribute looks significant. Hence, these attributes most likely contribute to liking to play video games. Unfortunately, independent of whether or not the individual likes to play video games, the survey data suggests there is an even split between finding video games educational or not. In light of this, our analysis focuses on whether designing a computer lab based on attributes of video games will be conducive to a more enjoyable lab session. The educational factor of video games will be satisfied through the fact that the lab session is for the purpose of education.

Evaluate Expected Grade Distribution to Target Grade Distribution?

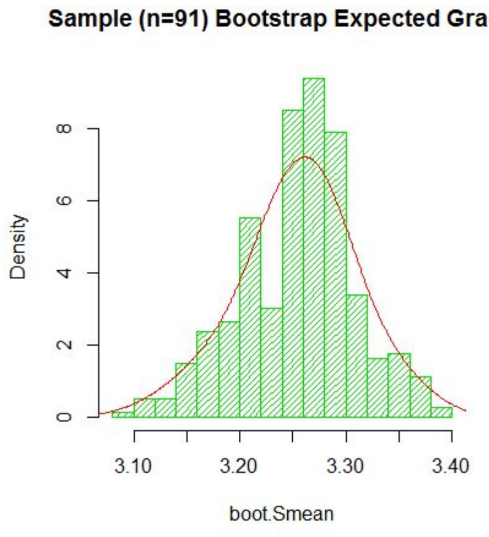
We want to evaluate the expected grades of the surveyed students. Specifically, how it matches up to the target grade distribution of 20% A's, 30% B's, 40% C's, and 10% D/F's. Also check how the distribution of the expected grades matches up to the target distribution when considering the 4 nonrespondents as failing students.

To begin, we look at the basic summary of the 'grade' value of our data:

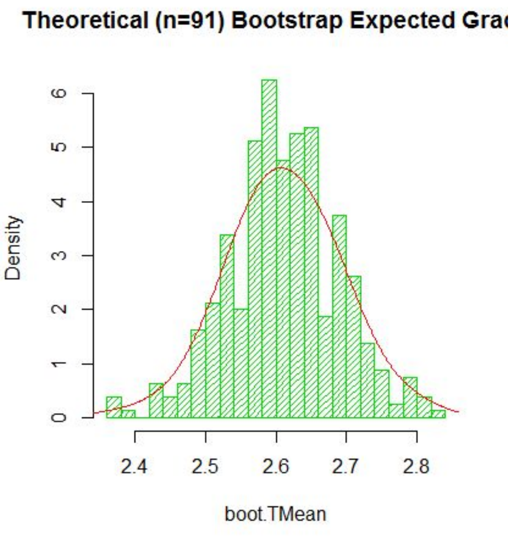
```
  Min.   1st Qu.  Median   Mean   3rd Qu.  Max.
 2.000    3.000    3.000   3.253   4.000   4.000
```

And so the average of our 'grade' data shows that the students as a sample expect a grade centered around **3.253**. When matched up against the target distributions average, **2.800**, we see that initially, the students in our sample collectively believe they will receive a higher grade than they actually will.

However, we would like to estimate how the class as a whole (**N = 314**) expects to perform in the class, rather than the sample that we have chosen (**n = 91**). Therefore, we bootstrap our given data to the N population size, making sure we pick our samples without replacement to mimic the probability model chosen for the survey.

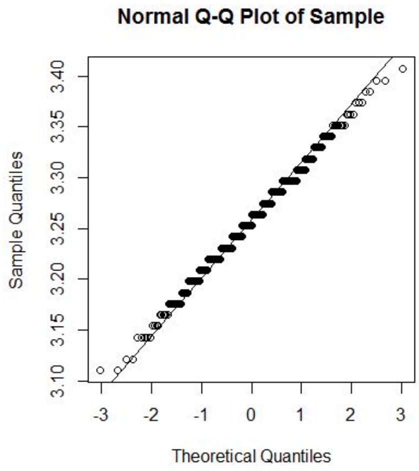


Plot 2.16: Sample Bootstrapped Population's Expected Grade with n = 91 expanded to N = 314 over 400 samples

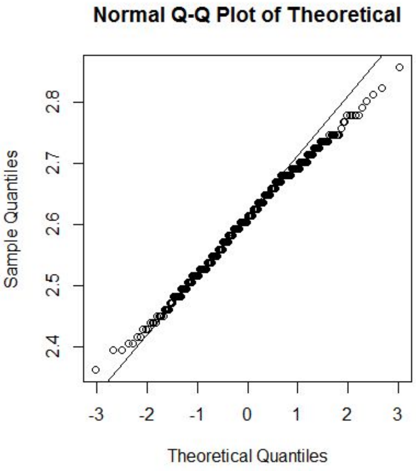


Plot 2.17: Theoretical/tTarget Bootstrapped Expected Grade With n= 10 (the target) expanded to N = 314 over 400 samples

What we have above is the result of our sample and target distribution (sample being the original dataset, and target being a generated dataset with the target distribution) of the sample means from each respective population, run 400 times. The bootstrapped population of the sample dataset shows a point estimate mean of **3.254**, whereas the bootstrapped population of the target/theoretical has a mean of **2.612**. While both distributions plotted are likely arise from a normally distributed dataset, the distribution of the sample dataset shows a slight skew to the right which can be an indicator of a confounder within the grade expected of the surveyed students.



Plot 2.18: Above is the Q-Q plot of the sample bootstrapped population



Plot 2.19: The Q-Q plot of the target bootstrapped population

The quantile-quantile plot to check normality of the sample dataset expected grade mean distribution shows likely normal distribution, however note the outliers which may be shifting our expected grade results by a confounding factor. The quantile-quantile plot of the target mean distribution has data that strays even further from our quantile-quantile line of normality.

To double-check if our dataset was actually constructed from a normally distributed sample, we performed the Shapiro-Wilk test:

Shapiro-Wilk normality test

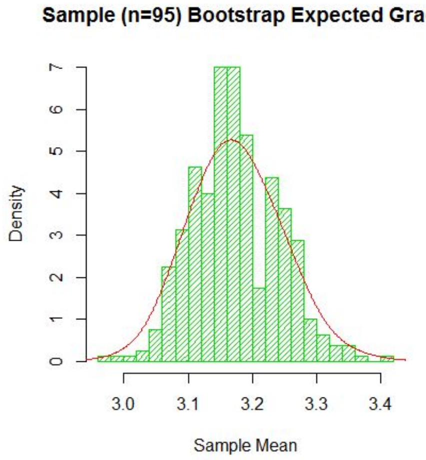
data: boot.Smean
W = 0.99258, **p-value = 0.04463**

Here, we have **p-value < 0.05**, which allows us to reject the null hypothesis of our expected grades from the sampled students being normally distributed.

Thus, it is likely that our students are overestimating the grades they will receive in the class, which could be the confounder affecting our bootstrapped population.

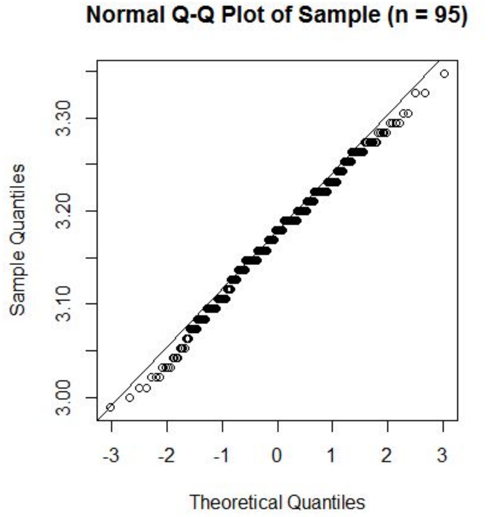
How Does the Distribution Change When Considering the Non-respondents as Failing Students?

In our survey methodology, it was noted that the survey size was to be originally the full 95 students in the class. However, **4** of the students were non-respondents and thus excluded from the survey. As such, we would like to now consider the expected grade for the full **95** students by adding in 4 entries with value ‘1’ for ‘grade’. This is equal to D/F range, which prior to this addition, interestingly enough, none of the students expected to fail.



Plot 2.20: Sample dataset bootstrapped population with n = 95, N = 314
over 400 samples

Notice, the mean of the new sample bootstrapped population’s expected mean grade now drops down to a point estimate mean of **3.173**. This is less than before, which indicates that the addition of **4** failing students brings our point estimate of the mean closer to the expected mean of a target distribution.



Plot 2.21:Q-Q plot of mean expected grade
with n = 95, the bootstrapped population run over 400 samples

While there is a dip near the bottom of the plot, we also see likely normal distribution of the mean expected grade. To double-check, we use the Shapiro-Wilk test of normality to test our null hypothesis:

```
> shapiro.test(boot.Smean)
```

Shapiro-Wilk normality test

data: boot.Smean
W = 0.99258, **p-value** = 0.04463

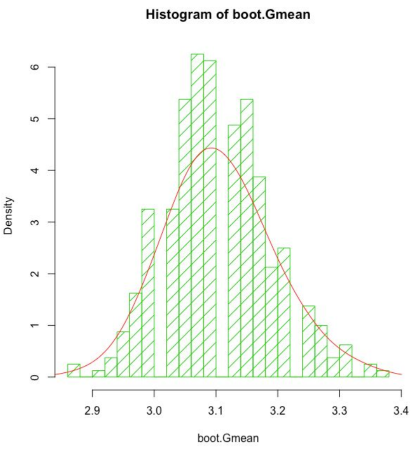
With **p < 0.05**, we can once again reject the null hypothesis: the mean expected grade of the bootstrapped population is not normally distributed.

This also follows because with 4/95 students expecting a D or lower, we do not have our ideal 10% of the population receiving D/F's, but rather only 4.2% receiving a failing grade. Overall, the mean is closer to the target but the picture does not change much: the students are still overestimating their final grades.

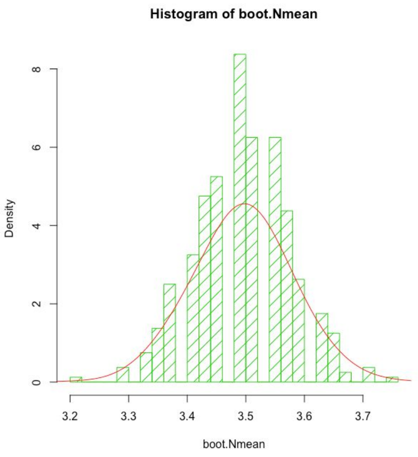
Do gamers or non-gamers have a better sense of their expected grades?

As a final addendum to this scenario, we wanted to check specifically how the expected grades distribution changes across both groups. From our survey methodology, we have the ‘frequency’ variable which seems like the best estimate of a gamer: how often they play, despite the changes in weekly workload. Of four possible responses, we picked the students with ‘1’ and ‘2’ responses under that category as “gamers” and the ‘3’ and ‘4’ responses as “non-gamers” (1 = daily, 2 = weekly, 3 = monthly, and 4 = semesterly).

Once the groups were chosen, the original **91** responses were grouped into **41** gamers and **37** non-gamers, with **13** unknown responses that were left out. With these remaining two groups (gamers and non-gamers), the sample set was bootstrapped to a population of **314** (N). Once the population was built for each group, we collected **400** samples and plotted the distribution of each of their mean point estimates, as was done previously in this scenario.

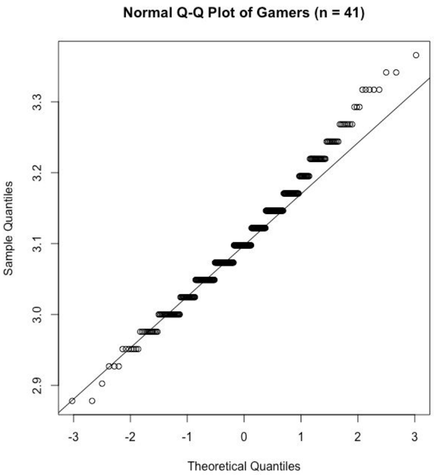


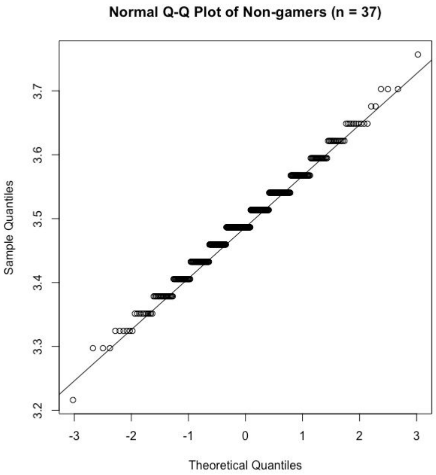
<u>Plot 2.22</u>: The gamer bootstrapped population sample means



Plot 2.23: The non-gamer bootstrapped population sample means

As shown above, we can see that the mean point estimate for each distribution shows that the gamers on average are closer to the real grade distribution they will receive. Whereas the non-gamers display a much higher tendency to overestimate their expected grades. The gamers distribution also shows a slight skew to the left.





Plot 2.24: Q-Q plot to check for normality of gamers

Plot 2.25: Q-Q plot to check for normality of non-gamers

Shapiro-Wilk normality test

data: boot.Gmean
W = 0.98767, **p-value** = 0.001822

Shapiro-Wilk normality test

data: boot.Nmean
W = 0.9852, **p-value** = 0.000414

With both **p-value < 0.05**, we reject both null hypothesis of our dataset being normally distributed. The gamers distribution of point estimate means from the bootstrapped population is far closer to the target distribution than that of the non-gamers.

Overall, while the students on the whole are likely overestimating their final grades for the course it is likely the gamers have a better understanding of their expected grade than the non-gamers. When considering the discussion lab, the amount of time gaming should be ample such that the non-gamers can get exposure to the benefits of gaming.

Conclusion

Given the analysis on the survey data, we have come to a few recommendations on how to design the computer lab. Our recommendations include:

- Students are affected by the stress of exams and will be less likely to play video games during that time because people didn’t spend that many hours playing video games on average. Therefore, we should have the lab meet for a short amount of time on the weeks there are exams OR have students play less in the lab that week.
- Students enjoy relaxing, feeling mastery, and education aspects of gaming, and those that don’t reported it cost too much or was a waste of time. Therefore, the lab will be better off focusing on one topic where the games approach it from different angles to provide the student with mastery of the subject.
- Based on the results of Scenario 5, the video game lab should not focus heavily on the educational aspect of the game, but rather aspects of video games that make them enjoyable. Specifically, it should include multiple difficulty levels to provide a challenge to the gamer students, as well as a plot-driven story to hide the educational aspects of the game.
- The lab should be designed for computer games which involve strategy relevant to the subject. Strategy games can be designed based on experimentation and optimization, which can pose as mental challenges for students.
- Since the games would have a constructive purpose, many students would feel that they are utilizing their time well and would also help them learn.

Theory

QQ Plots:

Histograms leave much to the interpretation of the viewer. A better graphical way in R to tell whether your data is distributed normally is to look at a quantile-quantile (QQ) plot. With this technique, you plot quantiles against each other. If you compare two samples, for example, you simply compare the quantiles of both samples. Or, to put it a bit differently, R does the following to construct a QQ plot:

- It sorts the data from both the samples.
- It plots these sorted values against each other.

If both samples don't contain the same number of values, R calculates extra values by interpolation for the smallest sample to create two samples of the same size.

In most cases, you don't want to compare two samples with each other, but compare a sample with a theoretical sample that comes from a certain distribution (for example, the normal distribution). To make a QQ plot this way, R has the special qqnorm() function. As the name implies, this function plots your sample against a normal distribution. You simply give the sample you want to plot as a first argument and add any graphical parameters you like. R then creates a sample with values coming from the standard normal distribution, or a normal distribution with a mean of zero and a standard deviation of one. With this second sample, R creates the QQ plot as explained before.

R also has a qqline() function, which adds a line to your normal QQ plot. This line makes it a lot easier to evaluate whether you see a clear deviation from normality. The closer all points lie to the line, the closer the distribution of your sample comes to the normal distribution. The qqline() function also takes the sample as an argument.

Regression Trees:

Regression trees are another form of nonlinear prediction trees that are useful for mixed data (i.e, data containing both numerical and categorical information). These are a form of recursive partitioning where the population is continuously split along the branches based on a classification of a data value and this subset of the population is analyzed making it easier to form theories on the actual distribution of complicated data. Each leaf node also includes the number of observations at that node and a proportion of the total observations at the branching point.

The plotcp() function is further analysis on the results obtained from the regression tree, based on the relative standard error, and the R-squared values which all give a visualisation of how good our predictive function is as a fit to the expected values.

Theory from class:

$Var(X_{i(j)}) = E(X_{i(j)} - \mu)^2 = \frac{1}{N}\Sigma_{i=1}^{N}(x_i - \mu)^2 = (sigma)^2$;

where $X_{i(j)}$: *when individual got into the sample from the population, a discrete variable*

$$E(Xbar) = 1/n(\sum_{i=1}^{n} E(X_{i(j)})) = 1/n \sum_{i=1}^{N} X_i P(I(j) = i) = \sum_{i=1}^{N} x_i(1/N) =$$

$$\mu X_{I(j)} = \begin{cases} 0, otherwise \\ Xi, If I(j) = i \end{cases} E(X_{I(g)})) = 0 * P(otherwise) + X_i * P(I(j) = i)$$

Confidence Interval:

These determine the quality of our estimator.

$$Var(Estimator) = Var(Xbar) = Var(1/n \sum_{j=1}^{n} X_{I(j)}) = Var(X_{I(j)}) =$$

$$E(X_{I(j)} - \mu)^2 = 1/N \sum_{i=1}^{N} (x_i - \mu)^2 = \sigma^2$$

$$(X_{I(j)} - \mu)^2 = \begin{cases} \mu^2, otherwise \\ (x_i - \mu)^2, \end{cases} E(x_{i(j)}) = \mu^2 * P(otherwise) + (X_i - \mu)^2 *$$

$$P(X_{I(j)=i}) = \sum_{i=1}^{N} (X_i - \mu)^2 * P((X_{I(j)} = i) + \mu^2 * P(otherwise) = \sum_{i=1}^{N} (X_i) *$$

$$P(I(j) = i)$$

Bootstrap:

When we use bootstrap, what we are trying to do is make an estimate of the population based solely off of the sample dataset. Specifically, we take the given dataset and repeat it until it adds up to the estimate of the population size, N. Once we have our “bootstrapped” population, we take a number of samples of size n, thus allowing us to resample the population.

Shapiro-Wilk Normality test:

This is a test for normality, by checking if the number of samples against the null hypothesis. If the test results in a p-level less than the chosen alpha level, we can reject the null hypothesis that the data is normally distributed.